

# Xử Lý Ngôn Ngữ Tự Nhiên

Annotation

Quy Nguyen

# Nội dung

1. Giới thiệu một số loại annotation
2. Vietnamese word segmentation

# Types of annotation

- ▶ Tách từ (word segmentation)
- ▶ Gán nhãn từ loại (Part-of-speech (POS) tagging)
- ▶ Gán nhãn thực thể tên riêng (Name Entity Recognition)
- ▶ Phân tích cú pháp dựa trên ngữ (constituent treebank)
- ▶ Phân tích cú pháp phụ thuộc (dependency treebank)
- ▶ Phân tích ngữ nghĩa (semantic roles)
- ▶ Text classification (Sentiment analysis)
- ▶ Question and answering
- ▶ Machine translation

# Text segmentation

- ▶ **Text segmentation** is the process of dividing written text into meaningful units, such as words or sentences.
  - ▶ Sentence segmentation
  - ▶ Word segmentation

## Sentence segmentation

- ▶ Ví dụ: Cho đoạn văn bản sau:

*"Cách bờ biển khoảng 10,5 km, TP Huế, tỉnh Thừa Thiên Huế, được xem là thành phố xanh với hơn 65.000 cây xanh đường phố, công viên. Bão Noul đã làm hơn 10.000 cây gãy đổ, nhiều nhất từ xưa đến nay. Một số công viên dọc bờ sông Hương như Thương Bạc, Phú Xuân, Bến Me, Kim Long..., nhiều cây cổ thụ nằm ngổn ngang trên mặt đất."*

Kết quả sau khi tách câu:

*<s> Cách bờ biển khoảng 10 km, TP Huế, tỉnh Thừa Thiên Huế, được xem là thành phố xanh với hơn 65.000 cây xanh đường phố, công viên. </s>*

*<s> Bão Noul đã làm hơn 10.000 cây gãy đổ, nhiều nhất từ xưa đến nay. </s>*

*<s> Một số công viên dọc bờ sông Hương như Thương Bạc, Phú Xuân, Bến Me, Kim Long..., nhiều cây cổ thụ nằm ngổn ngang trên mặt đất. </s>*

- ▶ **Thách thức: Nhập nhầm dấu câu**

# Word Segmentation

- ▶ **Tách từ** là thao tác chia văn bản thành những đơn vị từ.

- ▶ Ví dụ: Có câu sau:

<s>

*Hắn đã lập gia đình, nhưng cái gia đình ấy đã tan vỡ.*

</s>

Kết quả sau khi tách từ:

<s>

*Hắn đã lập gia\_đình , nhưng cái gia\_đình ấy đã tan\_vỡ .*

</s>

Hoặc có thể trình bày kết quả tách từ dưới dạng:

<s>

*Hắn/B đã/B lập/B gia/B đình/I ,/B nhưng/B cái/B gia/B  
đình/I ấy/B đã/B tan/B vỡ/I ./B*

</s>

# Part-of-speed tagging

- ▶ **"Part-of-speech (POS)**, also called grammatical tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context."<sup>1</sup>

- ▶ Ví dụ: Có câu sau:

*<s> Hấn đã lập gia đình, nhưng cái gia đình ấy đã tan vỡ.  
</s>*

Kết quả sau khi tách từ:

*<s> Hấn đã lập gia\_đình , nhưng cái gia\_đình ấy đã tan\_ vỡ  
. </s>*

Kết quả sau khi gán nhãn từ loại:

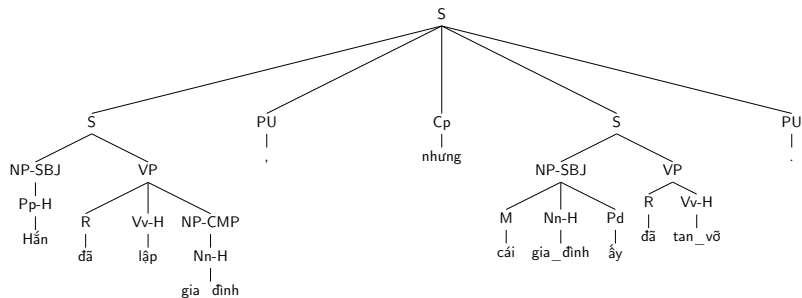
*<s> Hấn/Pp đã/R lập/Vv gia\_đình/Nn ,/PU nhưng/Cp  
cái/M gia\_đình/Nn ấy/Pd đã/R tan\_ vỡ/Vv ./PU </s>*

---

<sup>1</sup>[https://en.wikipedia.org/wiki/Part-of-speech\\_tagging](https://en.wikipedia.org/wiki/Part-of-speech_tagging) 

# Constituency treebank

## ► A Vietnamese constituency tree





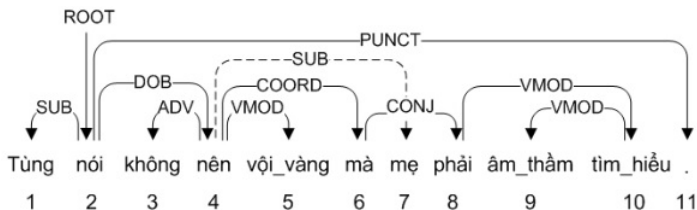
# Constituency treebank

- ▶ A Vietnamese constituency tree based on the Penn Treebank format

```
<s>
(S
  (S
    (NP-SBJ (Pp-H Hấn))
    (VP (R đã) (Vv-H lập)
      (NP-CMP (Nn-H gia_đình))))
  (PU ,)
  (Cp nhưng)
  (S
    (NP-SBJ (M cái) (Nn-H gia_đình) (Pd ấy))
    (VP (R đã) (Vv-H tan_vỡ)))
  (PU .))|
</s>
```

# Dependency treebank

Các nhãn trình bày sự phụ thuộc ngữ pháp giữa các từ trong câu




1	Tùng	_ N Np _	2	SUB	_ _
2	nói	_ V V _	0	ROOT	_ _
3	không	_ R R _	4	ADV	_ _
4	nên	_ V V _	2	DOB	_ _
5	vội_vàng	_ A A _	4	VMOD	_ _
6	mà	_ C C _	4	COORD	_ _
7	mẹ	_ N N _	4	SUB	_ _
8	phải	_ V V _	6	CONJ	_ _
9	âm_thầm	_ A A _	10	VMOD	_ _
10	tìm_hiểu	_ V V _	8	VMOD	_ _
11	.	_ . . _	2	PUNCT	_ _

# Named-entity recognition

- ▶ "Named-entity recognition (NER) (also known as (named) entity identification, entity chunking, and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc."<sup>2</sup>

---

<sup>2</sup>[https://en.wikipedia.org/wiki/Named-entity\\_recognition](https://en.wikipedia.org/wiki/Named-entity_recognition) 

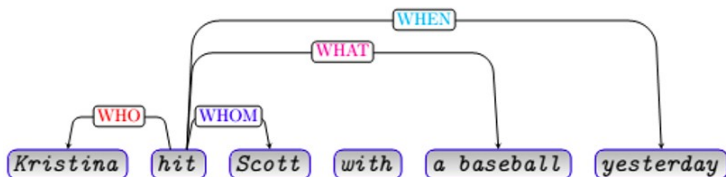
# Named-entity recognition

► Ví dụ:

Theo thống kê của <ENAMEX  
TYPE="ORGANIZATION">Sở y tế <ENAMEX  
TYPE="LOCATION">Hà Nội</ENAMEX></ENAMEX>,  
trong năm 2017 vừa qua có đến hơn 100.000 ca mắc bệnh sốt  
xuất huyết. Không chỉ sốt xuất huyết, rất nhiều dịch bệnh  
nguy hiểm khác bùng phát cũng do các côn trùng gây hại như  
bệnh dịch hạch, bệnh não... Theo Bách khoa tự điển  
<ENAMEX TYPE="LOCATION">Anh Quốc</ENAMEX>  
"Hầu hết các bệnh sốt nặng ở người đều do vi-rút truyền từ  
côn trùng gây ra".

# Semantic role labling

- ▶ Vai trò ngữ nghĩa là mối quan hệ giữa các thành phần cú pháp với vị ngữ (động từ chính)  
Ví dụ: *Agent (tác thể)*, *Patient (bị thể)*, *Instrument (công cụ)*, *Locative (vị trí)*, *Temporal (thời gian)*, *Manner (cách thức)*, *Cause (nguyên nhân)*, . . .



# Semantic role labling

- ▶ SRL rất hữu ích trong trong hỏi đáp. Nó giúp máy tính hiểu câu ở mức độ ngữ nghĩa nông, và có thể trả lời một số dạng câu hỏi

Ví dụ:

**Ai** đã đánh **Scott** bằng **một quả bóng chày**?

**Ai là người** đã bị **Kristina** đánh bằng **một quả bóng chày**?

**Kristina** đã đánh **Scott** bằng **cái gì**?

**Kristina** đã đánh **Scott** bằng **một quả bóng chày** **khi nào**?

# Time Annotation

- ▶ Biểu thức thời gian cho chúng ta biết:
  - ▶ Khi nào vấn đề xảy ra
  - ▶ Vấn đề xảy ra trong bao lâu
  - ▶ Vấn đề có xảy ra thường xuyên không

## Ví dụ:

- ▶ He wrapped up a **three-hour** meeting with the Iraqi president in Baghdad **today**.
- ▶ The king lived **4,000 years ago**.
- ▶ I'm a creature of **the 1960s, the days of free love**.

# Error Tagging

- ▶ Error tagging thường được làm cho các kho ngữ liệu dành cho người học
  - ▶ Cambridge Learner Corpora (CLC) and the Longman Learner's Corpus
- ▶ Các kiểu lỗi được sử dụng trong CLC
  - ▶ sử dụng sai từ
  - ▶ thiếu một cái gì đó
  - ▶ từ / cụm từ cần thay thế
  - ▶ từ / cụm từ không cần thiết

Ví dụ:

*My friend told me if I knew about Shakespeare. But, <TIP  
id=17-56 etype=24 tutor="I knew">I know</TIP>*



# Error Tagging

- Error tagging example:

```
<A>And, please describe this picture.</A>
<B>Describe? <F>Mhm</F>. <.></.> Maybe, <SC>this</SC>
<SC>toda</SC> <F>mm</F> it is a sunny day, and
<JP><F>unto</F></JP> in front of <SC?>hou</SC?>
<at odr="1" crr="a"></at> big house, <F>er</F> two
<n_inf odr="2" crr="housewives">housewives</n_inf>
<v_agr odr="3" crr="are">is</v_agr> talking
<prp_lxc2 odr="4" crr="to"></prp_lxc2> each other.
```

<at odr="1" crr="a"> (article, order of correction, correct form)

# Discourse and Pragmatic Annotation



# Vietnamese word segmentation

Vietnamese word types:

- ▶ **Single-syllabic words** are words that only have one free syllable<sup>3</sup>. A single-syllabic word can be a lexical word such as *quần*<sub>trousers</sub> and *hát*<sub>to sing</sub> or a function word such as *sẽ*<sub>will</sub> and *mà*<sub>that/which</sub>.
- ▶ **Coordinating compounds** are words that include two or more syllables, where the syllables can be single-syllabic words. However, the meanings of a coordinating compound are equally combined meanings of its components. For example, both *đất*<sub>land</sub> and *nước*<sub>water</sub> are single-syllabic words. However, if we treat *đất nước* as a coordinating compound, it means *country*.

---

<sup>3</sup>Free syllables are those having either a lexical or a functional meaning. A free syllable can stand alone as a word.

# Vietnamese word segmentation

Vietnamese word types:

- ▶ **Subordinate compound words** are words that include two or more syllables, where the syllables are combined according to a main-subordinate relationship. The main syllable is a word. The other syllables are not necessarily words. For example, *chân<sub>foot</sub>* and *vịt<sub>duck</sub>* are single-syllabic words. The combination of *chân* and *vịt* will create a subordinate compound word that means *presser foot (on a sewing machine)*. In the subordinate compound word *gầy guộc {skinny}*, *gầy<sub>thin/skinny</sub>* can stand alone as a single-syllabic word. However, *guộc* is a bound syllable<sup>4</sup> that does not have any meaning.
- ▶ **Reduplicative words** are constructed from the phonetic repetition phenomenon of syllables. Reduplicative words include two types, i.e., full word reduplication, such as *xa<sub>far</sub>* *xa<sub>far</sub> {in the distance}*, and partial reduplication, such as *long lanh {glistening}*.

---

<sup>4</sup>A bound syllable is a syllable that cannot stand alone as a single-syllabic word. A bound syllable does not necessarily have meaning. It always combines with other syllables or words to create a compound word.

# Vietnamese word segmentation

Vietnamese word types:

- ▶ **Reiterative forms** look like full reduplicative words. However, the reiteration form is contingently constructed by reiterating a word many times when we want to emphasize a large amount, a high frequency, etc. For example, *người người* {everybody}; *tối tối* {every night}.
- ▶ **Other multi-syllabic words** are constructed from syllables that do not have any meaning in Vietnamese. The syllables in a word do not have phonetic or semantic associations. The words can be originally Vietnamese, such as *bồ nông* {pelican}, or transcribed from Chinese such as *bù nhin* {puppet} or French such as *xà phòng* {soap}
- ▶ **Other types** that are considered in our WS guidelines are: proper names, names of laws, resolutions and agreements, telephone numbers, faxes, e-mail addresses, websites, nicknames, home addresses, idioms and locutions, numeral expressions, foreign words, abbreviations, and punctuations.

## Vietnamese word segmentation - How to treat bold text?

- ▶ Cô ấy mở cửa hàng **quần áo**.
- ▶ Bộ đồ này phối màu **quần áo** khác nhau.
- ▶ Khám phá nội dung mọi người đang **tìm kiếm**.
- ▶ Tôi mới được tận tay sờ vào chiếc **nồi đồng**.
- ▶ **Quần áo** bản thủ, nước da **đen đúa**.
- ▶ Kích thước của **cá heo** có thể từ 1,2 m.
- ▶ **Cá Lia Thia** ở nhiều vùng khác nhau mang tên gọi khác nhau.
- ▶ Tiêu chuẩn chức danh **nhà nghiên cứu** hạng III được quy định như thế nào?
- ▶ Tạo điều kiện cho **sinh viên** tiếp tục học tập.
- ▶ Mỹ đã bắt giữ Haizhou Hu, **nhà nghiên cứu** người Trung Quốc.
- ▶ Máy chạy **ầm ầm**.
- ▶ Đau đầu vì **quần quần áo áo**.
- ▶ Ở Thượng Hạ Cửu cũng tràn ngập toàn cửa hàng cửa hiệu, toàn **quần quần với áo áo**...

## Nine Vietnamese word segmentation rules

- ▶ If A and B have different meanings and the meaning of the combination form (A\_B) is different from the split form (A B), we select the form that has a more appropriate meaning for the context. as illustrated in rows 1 and 2 in Table 8.
- ▶ If A and B have different meanings and A\_B has the same meaning as A or B, A\_B is a compound word, as illustrated in row 3 of Table 8.
- ▶ If A and B have the same meaning, A\_B is treated as a compound word (row 4 in Table 8).
- ▶ If another syllable can be inserted between A and B, A and B are words (rows 5 and 6 in Table 8).
- ▶ If A or B (or both A and B) is bound syllable, A\_B is treated as a compound word. For example, we can not consider *đúa* in row 7 of Table 8 as a single word because it is a bound syllable. *đúa* does not have any meaning in Vietnamese. Hence, it is treated as a component of the compound word.

## Nine Vietnamese word segmentation rules

- ▶ For expressions composed by a categorization noun (A) and other words (B), if B indicates something different from what the expression indicates, A\_B is treated as a compound word. In contrast, if B has a similar meaning to A, A and B are treated as two words (rows 8 and 9 in Table 8).
- ▶ An expression composed by one or more Sino-Vietnamese syllables and an original Vietnamese word is not treated as a word when the Sino-Vietnamese syllables are the elements used to create similar words, such as antonyms and hyponyms. For example, *ngiên\_cứu*<sub>to research</sub> and *ngiên\_cứu viên* {researcher} in row 10 of Table 8 are similar words, in which the Sino-Vietnamese syllable *viên* plays the same role as the morpheme *-er/-or* in English. Expressions like *ngiên\_cứu viên* are not treated as words in our treebank. However, in cases as *sinh*<sub>to bear</sub> *viên* (row 11 of Table 8), we consider this expression as a word because *sinh*<sub>to bear</sub> and *sinh\_viên*<sub>student</sub> are not similar to each other.



# Nine Vietnamese word segmentation rules

- ▶ Special classifier nouns, e.g., *sự*–*ing*/–*ion*/–*ity*/..., *việc*–*ing*/–*ion*/–*ity*/..., and *nhà*–*er*/–*or* are treated as single words (row 12 in Table 8).
- ▶ If reduplicative words or reiteration forms are constituted by two syllables, they are treated as single words (row 13 in Table 8). In cases they have more than two syllables, if we can insert a comma or a conjunction between the syllables, we treat each syllable as a single word (rows 14 and 15 in Table 8). Otherwise, the whole expression is considered as a word (row 16 in Table 8).

# Vietnamese word segmentation

**Table 8** Examples of principles of word segmentation

No.	Expression (A B)	WS	Meaning
1	quần <sub>trousers</sub> áo <sub>shirt</sub>	One word	Clothes/clothing
2	quần <sub>trousers</sub> áo <sub>shirt</sub>	Two words	Trousers and shirt
3	ăn <sub>to eat</sub> nói <sub>to speak</sub>	One word	To speak
4	tìm <sub>to search</sub> kiếm <sub>to search</sub>	One word	To search
5	nồi <sub>pot</sub> đồng <sub>copper</sub>	Two words	Copper pot
6	nồi <sub>pot</sub> bằng <sub>by</sub> đồng <sub>copper</sub>	Three words	Copper pot
7	đen <sub>black</sub> đũa	One word	Black
8	cá <sub>fish</sub> heo <sub>pig</sub>	One word	Dolphin
9	cá <sub>fish</sub> lia_thia <sub>bettafish</sub>	Two words	Betta fish
10	ngiên_cứu <sub>to research</sub> viên_er	Two words	Researcher
11	sinh <sub>to bear</sub> viên_er	One words	Student
12	nhà_er nghiê_n_cứu <sub>to research</sub>	Two words	Researcher
13	ầm <sub>boom</sub> ầm <sub>boom</sub>	One words	Rumble
14	quần <sub>trousers</sub> quần <sub>trousers</sub> áo <sub>shirt</sub> áo <sub>shirt</sub>	Four words	Clothes/clothing
15	quần <sub>trousers</sub> quần <sub>trousers</sub> với <sub>with</sub> áo <sub>shirt</sub> áo <sub>shirt</sub>	Five words	Clothes/clothing
16	xa <sub>far</sub> lơ xa <sub>far</sub> lắ	One words	Very far